Тема 20

Държавен изпит

специалност

Приложна математика

Линеен регресионен анализ. Теорема на Гаус-Марков. Метод на най- малките квадрати

Анотация

Линеен регресионен анализ. Теорема на Гаус-Марков. Метод на най- малките квадрати

Въпросът може да се развие в общия случай или в случая с проста линейна регресия с Гаусова грешка. Използва се понятието максимално правдоподобие. Доказва се теоремата на Гаус – Марков, че максимално правдоподобната оценка и оценката на параметрите по метода на най-малките квадрати съвпадат. Привеждат се оценките за параметрите и тяхната дисперсия.

Линейния регресионен анализ (ЛРА) търси, описва и характеризира зависимости между величини, между които има корелация (зависимост) или поне се предполага, че има такава. В общия случай разполагаме с експериментални данни – измервания на определени величини, за които се предполага че са свързани. Целта обикновено е да се построи модел, чрез който поведението на някое от наблюдаваните явления да се предсказва въз основа на останалите. Явлението, за което търсим явен модел наричаме *отклик*, а тези, чрез които се моделира – *фактори*. За пример можем да се опитаме да предскажем какво ще е количеството риба в дадено изкуствено езеро въз основа на наблюдения от минали месеци/години на броя риби и нивата на концентрации на определени вещества във водата. В ЛРА данните са количествени.

Нека разглеждаме някаква (например физическа) система, в която провеждаме серия измервания на m+1 величини, една от които избираме за отклик – $\boldsymbol{y}$. Останалите са фактори и ги бележим с $\{x_i\}_{i=1}^{m}$. Провели сме серия експерименти. Данните от тях са

$$(1)\, y_i \leftrightarrow (x_{i1}, x_{i2}, \dots, x_{\Im})$$

Търсим линеен по коефициентите модел от вида

$$(2)\, y = a_1 x_1 + a_2 x_2 + \dots + a_m x_m$$

Ние предполагаме че такъв модел съществува, тоест, че природата на явленията е такава, че те са свързани по начина (2). Разбира се, рядко се намира модел от вида (2), в който като заместим данните (1) получаваме тъждество. Причината за това е че: първо, (2) е линеен по параметрите модел, а явлението, което искаме да опишем има нелинейна природа (в този случай можем да намерим само линейно приближение на истинския закон); второто, при измерванията както на факторите, така и на откликите се допускат технически и други грешки.

Нека въпреки това, от експерименталните данни (1) по някакъв начин сме предположили (2). Тогава имаме грешки от това предположение и те са

$$(3)\, \epsilon_i = y_i - a_1 x_{i1} - a_2 x_{i2} - \dots - a_m x_{\Im}$$

Ясно е, че ако тези грешки са малки, то моделът ни е добър и надежден за прогнози.

Възможни оценки за общата грешка на модела са например $\text{¿}\,\epsilon_i \vee \text{¿}, \sum_{i=1}^{n} \epsilon_i^2$ $\sum_{i=1}^{n} \epsilon_i, \sum_{i=1}^{n} \text{¿}$, от които вторите две са най- удачни. Същността на метода на най- малките квадрати, за който се отнася теоремата на Гаус- Марков, се състои в решаването на минимализационната задача

$$\min_{(a_1,\dots,a_m)} \sum_{i=1}^{n} \epsilon_i^2 = \min_{(a_1,\dots,a_m)} \sum_{i=1}^{n} \left(y_i - a_1 x_{i1} - a_2 x_{i2} - \dots - a_m x_{\Im}\right)^2$$

Ние ще изложим матричн подход към решаването на тази задача, но въпреки това, поради изпъкналостта на функцията $\sum_{i=1}^{n} \epsilon_i^2$ по $(a_1, ., a_m)$ имаме, че ако тя има локален екстремум, то той е и глобален. Системата, която се получава за намиране на критични точки е с m уравнения и m неизвестни и е линейна. Ако матрицата и е неособена, то тя има единствено решение.

Матричен запис

Тъждествата (3) могат да се запишат матрично по следния начин:

$$(4) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{pmatrix}_{n\times m} \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}_{m\times 1} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}$$

Вектор- стълбовете на матрицата M отговарят на факторите, а вектор- редовете на поредният експеримент. В по- съкратен запис (4) изглежда така:

$$(5)\, y = M.a + \epsilon$$

За да е коректен метода на най- малките квадрати е нужно вектор- стълбовете на M са са линейно независими. Това е еквивалентно да няма два идентични фактора, които взимаме в предвид или някой фактор да е резултат от друг. Преди да пристъпим към регресията е нужно да осигурим тази независимост на факторите.

Ние търсим оценка на вектора a, получен като статистика, състояща се от данните от (1). Да умножим (5) по M^T:

$$(6)\, M^T y = M^T M a + M^T \epsilon$$

От геометрични (хипер-геометрични) съображения, (които в някакъв смисъл са еквивалентни на съображението за екстремума, което направихме при обясняването на МНК чрез оптимизационната задача) изискваме $M^T \epsilon = 0.$ Тогава,

Ако умножим (6) по $(M^T M)^{-1}$ (която съществува ако стълбовете на M са линейно независими! да се има в предвид, че n>>m), получаваме оценката

$$(7)\quad \hat{a} = (M^T M)^{-1} (M^T y)$$

Най- важния факт в регресионния анализ е теоремата на Гаус-Марков, според която при определени условия, оценката (7) е максимално правдоподобна, т.е. е неизместена и с минимална дисперсия.

<u>Теорема</u> (Гаус- Марков)

Нека са дадени данните (1). Нека $\hat{a} = (M^T M)^{-1}(M^T y)$. Тогава ако $M^T \epsilon = 0$, $\sum_{i=1}^{n} \epsilon_i = 0$, $\epsilon_i \; N(0, \sigma^2)$, то измежду линейните оценки на а, $\hat{a}$ е максимално правдоподобна.

<u>Доказателство:</u>

Неизместеност:

$$E E\left(\left(M^T M\right)^{-1}\left(M^T y\right)\right) = \left(M^T M\right)^{-1} M^T E(y) = \left(M^T M\right)^{-1} M^T E(Ma+\epsilon) = ¿$$

$$¿\left(M^T M\right)^{-1} M^T (M E a + 0) = \left(\left(M^T M\right)^{-1}.\left(M^T M\right)\right). Ea = Ea = a$$

Минималност на дисперсията:

Нека $\tilde{a} = Cy$ е друга линейна неизместена оценка. Нека означим с $A = (M^T M)^{-1} M^T$ и нека $C = A + D.$ Но $\tilde{a}$ е неизместена, следователно

$$E = E(A+D)(Ma+\epsilon) = (A+D)(M E a + 0) = (A+D) Ma = AMa + DMa = ((M^T M)^{-1}.(M^T M)) a + DMa = a + DM.a$$

Следователно за да е неизместена $\tilde{a}$ е необходимо $DM = 0.$ Да пресметнем дисперсията на $\tilde{a}$:

$$Var = (A+D).Var(y).(A+D)^T = \sigma^2 (AA^T + AD^T + DA^T + DD^T) = ¿$$

$$AA^T + DD^T + (M^T M)^{-1}.(M^T D^T) + DM(MM^T)^{-1} = \sigma^2 (AA^T + DD^T) \geq \sigma^2 AA^T = Var$$

$$¿ \sigma^2 ¿$$

Последното е вярно, понеже DD^T е неотрицателно определена.

Забележка: В темата липсват някои неща от анотацията. Има я развита от доц. Матеев.

Литература

[1] Записки по статистика, Въндев

[2] Записки по ПС, спец ПМ, П. Матеев

Темата е разработена от Велико Дончев, уч. 2011/2012 г.